



# The Prokaryotic Genome Annotation Pipeline

A standard workflow for annotating prokaryotic genome assemblies

[https://www.ncbi.nlm.nih.gov/genome/annotation\\_prok/](https://www.ncbi.nlm.nih.gov/genome/annotation_prok/)

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

## Introduction

Genome annotation is a complex process that includes prediction of protein-coding genes and other functional genome units, i.e., structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, as well as transposons and other mobile elements. The NCBI Prokaryotic Genome Annotation Pipeline is designed to annotate bacterial and archaeal complete and draft genomes using a combination of *ab initio* gene prediction and homology based methods ([NBK147280](#)). This annotation pipeline, capable of processing a large data volume and currently being used by NCBI's RefSeq project, is offered as an annotation service to GenBank submitters. This service is unavailable for download or usage outside the NCBI submission portal due to its system and database dependencies.

## GenBank

You can request prokaryotic genome annotation when you submit your genome to GenBank. Note that NCBI offers two submission pathways (complete and WGS):

- **Complete genome:** A genome assembly can be submitted as a complete genome, as .sqn files through [Sequencing](#) [MacroSend](#) (see [Genome Submission Guide](#)) if it falls into either of the following cases:
  - ◊ You have sequenced the complete circular genome and there are no gaps.
  - ◊ You know the order and orientation of the contigs and were able to assemble your sequences, with Ns between the contigs, into a single scaffold representing the complete chromosome with no extra unplaced contigs.
- **Whole Genome Shotgun (WGS):** If the genome assembly is in multiple pieces that you were unable to assemble into a complete chromosome, then submit assembled scaffolds and/or contigs to our WGS database using the WGS submission portal. See the [WGS Genome submission](#) page for details.

Genome annotation results are made available to the submitter for review and modification, if needed, prior to finalizing the GenBank submission. In addition to providing protein and RNA feature annotation, meta-data is added to the record indicating the annotation method and software version used (details in the [software release note](#)). All **RefSeq** complete and draft bacterial and archaeal genomes, with the exception of RefSeq Prokaryotic [Reference Genomes](#), are annotated using NCBI's Prokaryotic Genome Annotation Pipeline (PGAP).

## Prokaryotic Genome Annotation Process

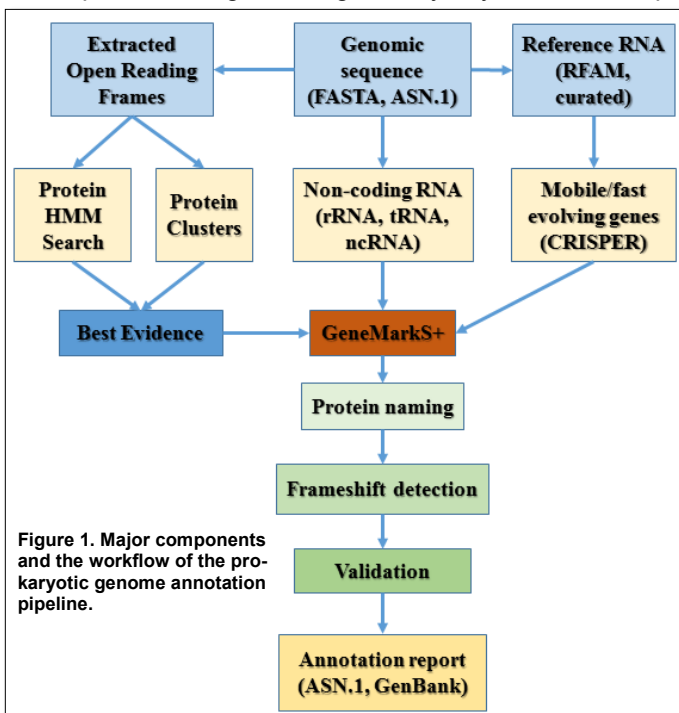
Gene prediction algorithms generally rely on statistical properties of the DNA sequence, and in some cases require a training set of genes typical for the organism. Some of the problems observed with *ab initio* prediction include:

- predicting coding genes in the region of non-coding RNA genes
- predicting two or more coding regions in the region of a single frameshifted gene
- giving a preference to a longer open reading frame and missing a conserved gene in the same region but in different frame
- incorrect prediction of atypical (phage-associated or horizontally transferred) genes, mobile elements and fast evolving systems (CRISPRs)

The NCBI pipeline combines an *ab initio* gene calling algorithm GeneMarkS+, developed in collaboration with the [GeneMark](#) team, with a similarity-based gene detection approach that minimizes many of these annotation problems and subsequent extensive post-processing steps. The pipeline currently predicts protein-coding genes, structural RNAs (5S, 16S, 23S), tRNAs, and small noncoding RNAs.

### Protein Alignments

The pipeline uses a pan-genome approach to collect a target set of proteins, consisting of universally expressed ribosomal proteins, clade-specific core proteins (conserved at the species or higher level), and curated bacteriophage and plasmid [protein clusters](#). Core proteins are identified through clustering at the clade level and must be found in the majority (~80%) of the available genomes. GeneMarkS+ uses alignments of the target protein data set to improve the consistency of genome annotation for closely related genomes.



## Protein Alignments (cont.)

[ProSplign](#), a frameshift-aware protein aligner. Complete gapless alignments with 100% identity to the target protein are accepted for the final annotation. Frameshifted alignments and partial alignments of good quality are passed to GeneMarkS+ for further refinement.

## Non-Coding RNA

- **Structural RNAs:** 5S, 16S, and 23S rRNAs are highly conserved in closely related prokaryotic species. The NCBI Reference Sequence Collection (RefSeq) contains a curated set of reference sequences for each subtype. The pipeline uses a BLASTn search against the reference set; 5S hits are further refined using [Cmsearch](#). Partial alignments that fall below 50% of the average length are dropped. Complete genomes failing this step are excluded from RefSeq.
- **tRNAscan:** To identify tRNA genes, the input genome sequence is split into ~200nt windows with an overlap of ~100nt, and passed through [tRNAscan-SE](#), which identifies 99–100% of transfer RNA genes in a DNA sequence while giving less than one false positive per 15 gigabases. It is currently one of the most powerful tRNA identification tools, with different parameter sets targeted for Archaea and Bacteria. Predictions with a tRNAscan-SE score below 20 are discarded.
- **Small ncRNA:** Small ncRNA prediction is a two-step process. First, the pipeline uses a BLASTn search against known FASTA sequences of selected [Rfam families](#). Subsequently, these regions are refined using Cmsearch with default parameters to produce the final annotation.

## Mobile/Fast Evolving Genes

- **Phages:** The annotation of phage related proteins is based on homology to a reference set of curated phage proteins. The phage reference data set comes from an independent effort of calculating and curating [protein clusters](#) from the complete bacteriophage genomes.
- **CRISPR:** CRISPRs (**C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats) are a family of DNA direct repeats of 20 to 40 nucleotides separated by unique sequences of similar length and are commonly found in prokaryotic genomes. These defense systems are encoded by operons that have an extraordinarily diverse architecture and a high rate of evolution for both the cas genes and the unique spacer content. The pipeline uses the CRISPR Recognition Tool (CRT) and PILER\_CR to identify and annotate CRISPRs.

## GeneMarkS+

Georgia Tech in collaboration with NCBI has developed a gene prediction program GeneMarkS+ that integrates information about protein alignments, frameshifted genes, non-coding RNA, and DNA statistical patterns typical for protein-coding and noncoding regions into gene predictions. GeneMarkS+ is an update of the published GeneMark algorithm ([Besemer, J. et al, 2001. Nuc Acids Res, 29: 2607-2618](#)).

## Frameshift Detection

Detecting frameshifts is a critical component of resolving ambiguities in automated annotation and assessing the quality of an assembly. The pipeline implements this as a two-step process. First, proteins from the target set are aligned to the genome with ProSplign that detects alignments with frameshifts, and aligned regions are passed to GeneMarkS+ and reported in the final output as pseudogenes. Second, newly predicted GeneMarkS+ genes are evaluated for potential frameshifts and aligned to the search set used for protein identification and naming. Several categories of potential frameshift, including a set of tandem models colocated within a threshold of 200nt and aligned to the same search protein, plus singleton partial alignments of sufficient quality, are identified as a potential frameshifted region. The region is extended 5' and 3' by a fixed (parameterized) amount. All candidate search proteins are then aligned to the region with ProSplign and evaluated for frameshifts. If at least four candidate search proteins align in a given region with a frameshift, the original models are replaced with a new gene feature with a pseudo qualifier covering the maximal extent of the aligned frameshifted proteins.

## Protein naming

The protein naming procedure makes use of curated protein names, domain content, curated HMM names, and BLAST search against a special naming BLAST database. The search set includes representatives from global protein clusters generated from all prokaryotic RefSeq genomes. Protein alignments are screened for quality (identity and symmetric overlap). A candidate protein is assigned to an identification cluster if there are a sufficient number of high quality search set proteins that point consistently to the same cluster.

## Annotation results

The annotation pipeline produces ASN.1 files (\*.sqn) ready for GenBank submission, convertible to traditional GenBank flat file for manual review. A summary report is generated to report the total number of predictions by each feature type.